Journal of
Next Generation Sequencing & Applications

Munoz-Minjares and Shmaliy, Next Generat Sequenc & Applic 2016, 3:3
DOI: 10.4172/2469-9853.1000141

Short Communication Open Access

The Role of Optimal Detection of CNAs and Error Analysis Using Next Generation Sequencing

Jorge Munoz-Minjares and Yuriy S Shmaliy*

Department of Electronics Engineering, Guanajuato University, Salamanca, Mexico

***Corresponding author:** Yuriy S Shmaliy, Department of Electronics Engineering, Guanajuato University, DICIS, Ctra. Salamanca-Valle, Palo Blanco, Salamanca 36855, Mexico, Tel: +524777145859; E-mail: shmaliy@ugto.mx

Rec date: Nov 01, 2016; **Acc date: Dec 29**, 2016; **Pub date:** Dec 31, 2016

Short Communication

Next generation sequencing is a concept that involves several different modern technologies to sequence large stretches of deoxyribonucleic acid (DNA) with wide advantage in throughput and scale compared to antiques sequencing technologies.

Main benefits of the next generation sequencing (NGS) are recognized as follows [1]:

- High-resolution, which allows obtaining a base-by-base view of a gene, exome, or genome.
- Quantitative measurements based on signal intensity.
- Detection all types of genomic alterations, including single nucleotide variants, insertions and deletions, copy number changes, and chromosomal aberrations.
- High throughput and flexibility in scaling and sequencing multiple samples simultaneously.

These characteristics have brought revolutionary advances to genetic field and resulted in the development of a wide variety of methods, which allowed researchers to ask virtually any question related to the genome, transcriptome, or epigenome of any organism. But in spite of advantages of the NGS technology, an important flaw inherent to earlier developed methods still remains. The NGS data are contaminated by intensive noise that requires using efficient methods of statistical signal processing and bioinformatics to eliminate undesirable information, so that NGS data could finally be used by researchers to make a clinical interpretation.

The methods, which are developed to study the heritable or acquired alterations in the DNA are called genomics. Some examples of this technology are the whole-genome Sequencing, exome sequencing, *de novo* sequencing and targeted sequencing. The whole genome association studies aim to identify the genetic basis of traits and disease susceptibilities using SNP microarrays that capture most of the common genetic variation in the human population [2]. In [3], one can find a report related to the genome-wide copy number variations, single nucleotide mutations and DNA methylation findings that might be specific in African-American colorectal cancer patients.

Single-nucleotide polymorphism (SNP) arrays simultaneously define the copy-number changes and allelic imbalances (including LOH) occurring in a tumor, at high resolution and throughout the whole genome [4]. SNP arrays are presently one of the most efficient technologies for the identification of copy number alterations [5].

Accurate detection of copy number alterations-aberrations (CNAs) with high accuracy is one of the main objectives of the technologies based on the NGS. The CNAs represent changes in the chromosomal structure, which result in gains or losses in copies of sections of DNA and are often associated with many types of cancer [6]. What many issues face in practice of the CNA detection is that an intensive noise does not allow for an accurate detection of the breakpoints and precise estimation of the segmental levels. In fact, accuracy in the estimates of the segmental levels is strongly affected by the segmental noise [7], which is widely recognized to be white Gaussian [8]. In view of the a fore mentioned issues, many approaches have been developed during decades to provide efficient NGS data denoising while preserving edges in the CNA structures.

Our principal research is based on the study of the statistical characteristics of data provided using the NGS technology and on the development of the confidence probabilistic masks, which allow improving the CNA estimates provided by diverse estimators. Specifically, a great attention is paid to the probabilistic analysis of jitter in the breakpoints of the CNA profiles [9] referring to the following widely recognized properties of the CNA function [8]:

- It is piecewise constant (PWC) and sparse with a small number of alterations on a long base-pair length.
- Its constant values are integer, although this property is not survived in the log R Ratio.
- The measurement noise in the log R Ratio is highly intensive and can be modelled as additive white Gaussian.

Based on our recent studies of the confidence limits for the stepwise signals measured in white Gaussian noise, we have developed an efficient algorithm for computing the confidence upper and lower boundary masks to guarantee an existence of genomic changes in certain regions with required probability. The masks are formed based on the segmental noise Gaussian distribution and jitter distribution in the breakpoints. We have shown that, in the first order approximation, the jitter in the breakpoints is distributed with the discrete skew Laplace law, which gives satisfactory results for the signal-to-noise ratio (SNR) exceeding unity. Because lower SNR levels are also of high importance, we have proposed several approximations for the jitter distribution with low and highly low SNR levels. The approximations we have developed are based on the modified Bessel function of the second kind and zeroth order. We have also modified the discrete skew Laplace law to have the noise variance dependent on the displacement with respect to the candidate breakpoint. We suggest combining these masks with estimates in order to give medical experts more information about the true CNA structures. Extensive investigations of the NGS-based measurement data provided using the confidence masks ensure that there is always a probability that some changes detected by an estimator do not exist and can be ignored [10], while some others exist in certain regions and not at exact points predicted by an estimator [11].

In conclusion, the modern NGS-based technologies have provided a high resolution in sequencing that, however, caused an essential increase in the computational complexity of data analysis. On the other hand, while processing data obtained by the NGS, an estimator discovers more subtle chromosomal effects than before that cause another problem. In the presence of intensive noise, not all of the detected CNAs may exist with a probability required. Moreover, many detected segments and breakpoints may exist in certain regions specified by the confidence probability rather than at exact points suggested by an estimator. For this reason, methods of statistical signal processing and bioinformatics will play a more essential role with further development of the NGS-based technologies. This relates to optimal and robust algorithms supplied with methods of error analysis. Final results produced by these techniques must be tested and interpreted by experienced genetic experts.

References

1. Illumina (2014) An introduction to next-generation sequencing technology.
2. Pauline C Ng, Kirkness EF (2010) Whole Genome Sequencing. Genetic Variation, Volume 628 of the series methods in molecular biology, Chapter 12. pp. 215-226.
3. Brim H, Ashktorab H (2016) Genomics of colorectal cancer in African Americans. Next Generat Sequenc & Applic 3: 133.
4. Dutt A, Beroukhim R (2017) Single nucleotide polymorphism array analysis of cancer. Curr Opin Oncol 19: 43-49.
5. Engle LJ, Simpson CL, Landers JE (2006) Using high-throughput SNP technologies to study cancer. Oncogene 25: 1594-1601.
6. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444(7118): 444–454.
7. Munoz-Minjares J, Shmaliy YS, Cabal AJ (2014) Noise studies in measurements and estimates of stepwise changes in genome DNA chromosomal structures. Adv Appl Pure Math 212-221.
8. Pique-Regi R, Ortega A, Tewfik A, Asgharzadeh S (2012) Detection changes in the DNA copy number. IEEE Signal Processing Mgn 29: 98-107.
9. Munoz-Minjares J, Shmaliy YS (2013) Approximate jitter probability in the breakpoints of genome copy number variations. Electrical Eng computing science and autom control (CCE), Mexico City, Mexico 128-131.
10. Munoz-Minjares J, Cabal-Aragon J, Shmaliy YS (2014) Confidence masks for genome DNA copy number variations in applications to HR–CGH array measurements. Biomed Signal Process Contr 13: 337-344.
11. Munoz-Minjares J, Shmaliy YS (2017) Improving estimates of the breakpoints in genome copy number alteration profiles with confidence masks. Biomed Signal Process Contr 10: 238-248.